THE DESCRIPTION OF SPOKEN LANGUAGE

J.Bruce Millar

Computer Sciences Laboratory
Research School of Physical Sciences and Engineering
Australian National University

ABSTRACT

This paper outlines the importance of a principled description of spoken language for the future of speech and language technology and proposes an initial structure for the description of spoken language which will draw on the perspectives of professionals from a range of disciplines. A principled description of spoken language is deemed important for the development of speech and language technology for two major reasons: (1) for the efficient use of resources, and (2) for effective assessment of the technology. The implementation of such a spoken language description is described in the context of the ANDOSL (Australian National Spoken Language Database) project. This implementation is seen to be conveniently structured as a hierarchy of levels containing mostly technical but also political and even legal material. The relationship between this structure and other structures in use is discussed.

INTRODUCTION

The task of describing spoken language data in an exhaustive fashion is daunting for several reasons. These include its complexity, its sources of unexplained variance, and the varied perspectives of the different spoken language professional disciplines. However, in the context of defining the structure of a spoken language database, it is important that more complete, yet extensible, descriptive models are developed.

Descriptive information has normally been held in a header block attached to the data, and, significantly, in a variety of documentation forms often not machine-readable, or even in a structured form. For reasons of economy and simplicity the data stored in a structured header is often minimised to include only those data essential for a range of signal processing tasks. Further, what is deemed useful and the form in which it is stored varies from system to system. The passing of data between systems by conversion routines which transform one form of header into another can very soon result in loss of most descriptive data as the thread of common descriptors across systems is broken.

The thesis of this paper is that descriptive information needs to be formalised and placed in a recognised standard structure. This standard descriptive structure may operate in parallel with or independently of existing header structures. Mappings can be simply generated between a rich descriptive structure and any less rich but convenient structure that is desired. What will have been gained is a common reference frame and a machine-readable form in which descriptive data may be preserved.

The challenge is that the sheer amount of potential descriptors is very large. What is proposed is a structure that has a place for any descriptor and a suitable form in which it can be expressed. This provides an "address" at which this information about the data can be found if it is available. Initially there should be an agreed set of descriptors and a way in which "new" descriptors can first be used in a non-interfering way, and then registered for use within the standard.

Such a principled description of spoken language is deemed important for the development of speech and language technology for two major reasons: It will promote the efficient use of resources by increasing multiple use of data and it will enable more effective assessment of speech technology as data on which systems are evaluated can itself be quantitatively assessed.

EFFICIENT USE OF RESOURCES

Precisely because spoken language is such a complex entity, it is capable of being described in a number of different ways which can tend towards being mutually exclusive. The worlds of the socio-linguist, the telecommunications engineer, and the phonetician can be almost totally disjoint. The fact that they are not is due to the fact that they all rely on the existence of spoken language. That common ground can be very slight as their modes of description of the phenomenon that unites them diverge at a very early stage in their treatment of speech. A consequence of this divergence is that, even within, and certainly between, such disciplines there exists a wide degree of variance in the way speech is described.

A consequence of this variance is that standards for spoken language description have been regarded as either pointless or perhaps just too hard to implement. Much duplication of effort has resulted from this attitude. Data is collected for new projects because what is available lacks the necessary descriptors to give confidence to the new user that it meets their requirements.

It is perfectly clear that no data is likely to be collected that will be described to a level of detail that would be regarded as comprehensive by all potential users of the data. It is possible however to move in the direction of comprehensive description so as to create a significant overlap of potential usefulness for the data so described.

A major practical benefit of this approach is that data gathering is approached from the perspective of its multiple use. Multiple use implies enhanced efficiency by the sharing of the cost of production of a data corpus across more than one project. Data description standards of themselves will not achieve efficiency but they will promote it. Completing a questionnaire about the data will alert the collector to the additional information that is deemed important by colleagues within the broad field of speech communication research, and to certain standards which would enhance further the re-useability of the data.

ASSESSMENT OF TECHNOLOGY

The assessment of speech technology is an issue of at least the complexity of the nature of spoken language. Over the years many very superficial claims have been made for speech technology performance. This problem is being addressed by a series of international meetings which commenced in the Netherlands in 1989, and have continued in Japan, Italy, and Canada in subsequent years. An international coordinating committee and three working groups, on speech recognition assessment, speech synthesis assessment, and speech corpora, have been established. Speech data which can be accepted and used by researchers and developers around the world is a major goal.

A DATA DESCRIPTION PROPOSAL

An initial taxonomy has been developed for the description of the proposed Australian National Database of Spoken Language (ANDOSL). The broad outline of this approach and the necessary components of a descriptive scheme have been described by Millar et al (1990a; 1990b) and extended in the area of speaker characteristics by Millar (1991).

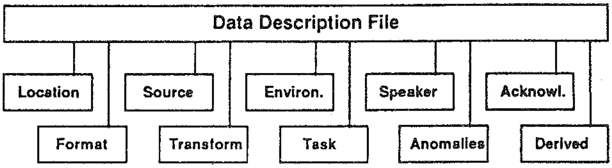


Figure 1

It is proposed for each data file that there be an associated "data description file" in which basic descriptors and pointers to more complex and maybe specialist descriptors have a common and agreed form. Therefore this prototype taxonomy has a hierarchical structure. At its top level ten basic categories of description are defined (see Figure 1). These are (1) the location of the data file, (2) the format of the data file, (3) the source of the data file, (4) the transformations applied to the source to derive the data, (5) the acoustic environment of the recording, (6) the nature of the speaking task given to the speaker, (7) the characteristics of the speaker, (8) any anomalies that occurred in the recording process, (9) any acknowledgements or restrictions that apply to the use of the data, and (10) a directory of "derived" information pertaining to the data.

It can be seen that this top level of the descriptive data structure is an amalgam of mostly technical, but also political and even legal issues all of which are important for a potential user of spoken language data to assess. There is considerable structure below this top level of description especially in the categories of speaker characteristics and "derived" data. Subsequent levels encapsulate data structures optimised to contain the information pointed to by the top level.

Certain of these top-level categories contain information which is intimately related to the data file being described and is not relevant to any other data file stored in the database. Other categories such as environment, speaker, task and acknowledgements contain information which has relevance to many data files. For the sake of economy, the former type of descriptors are stored within the DDF file itself, and the latter type are held in independent files which are referenced by pointers from within the DDF.

LOCATION OF DATA FILE

```
Location

DBN: <database_name> % ANDOSL
PCN: <primary_corpus_name> % ACRB-82
CCN: <current_corpus_name> % PASSAGES

NAD: <network_address> % "andosl.anu.edu.au"
IPN: <IP_number_of_site> % "130.56.6.117"
VOL: <volume_name> % "reo0"
DIR: <directory_path> % "aug92"
DFN: <data_file_name> % "TR123456"
```

Figure 2

Location Description

It is assumed that data may be stored in a distributed fashion. This may simply be across a set of removable packs of magnetic media, or it may be across a computer network. It is essential that an adequate "address" of each data item is contained in the descriptive database so that once it has been selected it may be accessed.

FORMAT OF DATA FILE

```
Format

LOH: <length_of_header> % number of bytes
TOH: <header_termination> % character string
HTV: <header_type_version> % file defining header

NCH: <number_of_channels> % Interleaving factor
DFO: <data_format> % "ASCII","BINARY","BLOCK"
ASF: <ascii_format> % "integer_per_line"
BNF: <binary_format> % "offset_bin","2s_comp"
NBS: <number_bytes_sample> % "2"
SBO: <sample_byte_order> % "low/high"
BLF: <block_format> % "ILS_ANAL","MSCC_12"
```

Figure 3

Format Description

The signal data may be held in a number of different formats. The format description allows for headers to be present, for multiple channels, and for ASCII, binary, or block formatted data.

SOURCE OF DATA

The data within one data file may represent a complete "take" of a speech task performed by the speaker, but in many cases it will be incomplete due to segmentation or preprocessing, such as downsampling,

```
Source
PRF: <primary_represent_file> % "wd804:
STA: <start_wrt_PRF> % start sample in PRF
END: <end_wrt_PRF> % end sample wrt PRF
RID: <representation_id> % "word","breath_group"

RED: <recording_date> % date in [YEARMMDD]
RET: <recording_time> % time format [HHMM]
REP: <recording_place> % recording facility

SPS: <samples_per_second> % "20000"
BPS: <bits_per_sample> % "16"
DQF: <disp_quantum_function> % "linear","logarithmic"
MAV: <maximum_decimal_value> % "32768"
MIV: <minimum_decimal_value> % "-32768"
```

Figure 4

Source Description

for economy of storage. Incomplete "takes" imply missing information that is potentially useful to the future user of the data. Clear reference to the original will enable missing information to be found. The time-place of recording, the digitisation parameters, and the portion of the original are also encoded.

TRANSFORMATIONS APPLIED

```
Transform
TIF: <filename>
--------------------
TCn: <transform_process_code> % eg "DOWNSAMP","LPC"
Tnm: <transform_process_param> % process parameters
```

Figure 5

The processes used in deriving the data from its primary recorded form are essential information in characterising the data fully (Millar, 1972). The presence of this field in the DDF pointing to a "transform identification file" (TIF) frees the description to apply to a wide variety of "speech data types" including such variants as Mel-scaled cepstral coefficients whose catalogue of transformations stored in a named "TIF" file would be considerable.

RECORDING ENVIRONMENT

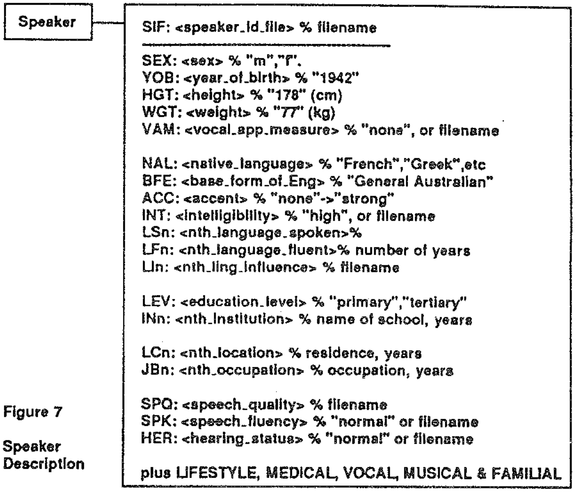
```
Environ.
EIF: <environment_id_file>% file containing rest
--------------------
MNM: <microphone_name> % identifier
MTY: <microphone_type> % e.g. "condenser"
MSP: <microphone_specs> % filename

MDT: <microphone_distance>% distance in mm
MEL: <microphone_elev> % elevation (degrees)
MAZ: <microphone_azimuth> % azimuth (degrees)
MOR: <microphone_orient> % orientation (degrees)

SPR: <speaker_posture> % "sitting","standing"
AOH: <azimuth_of_head> % azimuth (degrees)
EOH: <elevation_of_head> % elevation (degrees)
RSN: <reflective_surface> % e.g. "wood","paper"
RSD: <reflective_surf_dis>% from mouth (mm)
RSO: <reflective_surf_ori>% orientation (degrees)

ASN: <ambient_sound_nat> % eg "impulse","speech"
ASB: <ambient_sound_band> % Hertz range
ASL: <ambient_sound_level>% dB (SPL)

PRE: <passive_rec_environ>% eg "office","car"
...
AAF: <anti_alias_filter> % eg "butterworth-8"
...
CHP: <channel_processing> % filename
...
```

Figure 6

Environment Description

There are several features of the environment in which the recording was made that influence the form of the speech. These can include psychological factors prompted by the unnaturalness of the environment, but also many physical factors related to the relative positions of transducers and speakers, or to the acoustic characteristics of the environment. The latter can be both active (intrusive sounds) or passive (the acoustic absorption of surfaces).

SPEAKER CHARACTERISTICS



Speaker

```
SIF: <speaker_id_file> % filename

SEX: <sex> % "m","f".
YOB: <year_of_birth> % "1942"
HGT: <height> % "178" (cm)
WGT: <weight> % "77" (kg)
VAM: <vocal_app_measure> % "none", or filename

NAL: <native_language> % "French","Greek",etc
BFE: <base_form_of_Eng> % "General Australian"
ACC: <accent> % "none"->"strong"
INT: <intelligiblity> % "high", or filename
LSn: <nth_language_spoken>%
LFn: <nth_language_fluent>% number of years
LIn: <nth_ling_influence> % filename

LEV: <education_level> % "primary","tertiary"
INn: <nth_institution> % name of school, years

LCn: <nth_location> % residence, years
JBn: <nth_occupation> % occupation, years

SPQ: <speech_quality> % filename
SPK: <speech_fluency> % "normal" or filename
HER: <hearing_status> % "normal" or filename

plus LIFESTYLE, MEDICAL, VOCAL, MUSICAL & FAMILIAL
```

Figure 7

Speaker Description

Speaker attributes which can influence the form of their speech are numerous. They include physical, linguistic, educational, geographical, occupational, voice quality, speech or hearing pathology, medical, lifestyle, vocal training or abuse, musical, and familial factors. A base set of variables is proposed to encode the main sub-factors within these major groupings. The full set is contained in a hierarchy of files of which the root is the "SIF" file pointed to by the DDF.

SPEECH TASK

The nature of the speech task given to the speaker, including reading or prompting material, is of considerable importance for correct interpretation of the acoustic data. The task may be very artificial yielding very stylised speech including such phenomena as "list intonation" or "hesitation". A more natural task will produce speech which is probably articulated with less precision but without artifical stylisations. The substructure of the "TIF" file is still to be defined.

ANOMALIES

A list of any known anomalies experienced in the data preparation, is a very important contributor to the interpretation of the data. A series of mispronunciations, a bout of coughing, a headache or a very tense day can unsettle a speaker and create anomalous patterns in their speech. The anomalies text file can be used to flag general factors such as health and tension and specific non-standard epochs in the data collection period.

RESTRICTIONS AND ACKNOWLEDGEMENTS

A text file of appropriate acknowledgements which should be made by users of the data and of any restrictions or costs which apply to its use, is important as it represents the legal and/or moral responsibility of the user of the data to acknowledge the intellectual property inherent in the data.

DERIVED FILES DIRECTORY

In addition to these basic "extrinsic" attributes of the data, there is provision for a set of "intrinsic" qualities which have been derived from the data itself. Such derivations are often laborious, involve special skills, or are derived from a class of data items of which the current sample is but a part. Their extent cannot be defined as they are limited only by the interests of the population of users. Such sets of derived files are likely to apply to a range of data and are therefore listed in a named "DFD" file. This list may include several generic types of information; all forms of acoustic features, phonetic and linguistic annotation, distributional characteristics such as the percentile value of speaking rate, as well as simple signal characteristics such as signal-to-noise ratio.

CONVERSION STRATEGIES

A strategy has been developed for the interconversion of descriptive data between this potentially very rich scheme and less rich but entrenched descriptive formats. This involves "import" and "export" drivers for the proposed scheme which map its categories onto the restricted category sets of such header systems as those of the Interactive Laboratory System (ILS), the National Institute for Standards in Technology (NIST), the Entropic Signal Processing System (ESPS), and the Speech Assessment Methodologies (SAM) project. Prototype software for this strategy has been developed. It allows basic storage of the data description to be in the proposed rich format but the use of other popular formats without losing data in multiple conversions between less-rich descriptive schemes.

THE WAY FORWARD

This paper has outlined the basic structure and given a major insight into the content of the upper levels of a proposed "standard" for spoken language data description. Ongoing work involves the refinement of the upper levels, their interface with lower levels, and the definition of the lower levels which tackle such issues as accented speakers, pathological speech, the description of the skills of the annotator, and the normalisation indicators for a range of subjective judgements.

ACKNOWLEDGEMENTS

The author is most grateful to the following people who have assisted his thinking on the issues presented here. John Ingram, Jenni Oates, Lynda Penny, Julie Vonwiller have been willing discussants of issues relating to the interface between "normal" and accented or pathological speech. Arthur McGuffin and Owen Genat have written a prototype program to convert between this structure and several other structures in common use. Mugambi Kiruki has investigated the role of an object-oriented structure to encode this data structure.

REFERENCES

Millar,J.B. (1972) *An interactive speech processing system using a large computer* Int. J. Man-machine studies, Vol.4, pp.285-317.

Millar,J.B., Dermody,P., Harrington,J.M., Vonwiller,J.P. (1990a) *A national spoken language database: concept, design, and implementation* In "Proceedings of International Conference on Spoken Language Processing (ICSLP-90)", Kobe, Japan, 18-22 November pp.1281-1284.

Millar,J.B., Dermody,P., Harrington,J.M., Vonwiller,J.P. (1990b) *A national cluster of spoken language databases for Australia* In "Proceedings of Third Australian International Conference on Speech Science and Technology", Melbourne, 27-29 November, pp.466-471.

Millar,J.B. (1991) *Knowledge of speaker characteristics: its benefits and quantitative description* In "Proceedings of XII International Congress of Phonetic Sciences", Aix-en-Provence, 91-24 August, pp.538-541.